Letter to the Editor

Improving the Estimation of Bacterial Allele Frequencies

Eric C. Anderson* and Paul A. Scheet†

*Interdisciplinary Program in Quantitative Ecology and Resource Management, University of Washington, Seattle, Washington 98195 and †Department of Statistics, University of Washington, Seattle, Washington 98195

Manuscript received January 9, 2001
Accepted for publication April 30, 2001

RANNALA et al. (2000) develop a useful Poisson process model for estimating allele frequencies in bacterial populations using presence/absence data on infected hosts. However, they also show that the estimators they derive suffer from bias. In this letter we derive different estimators from the same model and show they derive in both of the following sections are for allele frequencies \( p \) and also simplifies the estimation of \( p \) and \( \lambda \) as described in estimation from the whole sample of hosts. The MLEs derived in both of the following sections are shown to be less biased than the estimators given in RANNALA et al. (2000).

CONSTRUCTED OPTIMIZATION

Any values of \( p \) and \( \lambda \) that maximize (1) while satisfying the constraint \( \sum_{j=1}^{k} p_j = 1 \) will also be maximizers of the equation

\[
\mathcal{L} = -n \log(1 - e^{-\lambda}) + \sum_{j=1}^{k} \left( z_j \log(1 - e^{-\lambda p_j}) - (n - z_j)\lambda p_j \right),
\]

where \( \mu \) is an arbitrary nonzero constant (\( \mu \) is called the “LaGrange multiplier” in this context). The MLEs may be found by setting the first partial derivatives of (2) with respect to \( p \) and \( \lambda \) equal to 0 and solving for \( p \), \( \lambda \), and \( \mu \). Thus we have

\[
0 = \frac{\partial \mathcal{L}}{\partial p_j} = \frac{\lambda z_j e^{-\lambda p_j}}{1 - e^{-\lambda p_j}} - (n - z_j)\lambda + \mu, \quad j = 1, \ldots, k
\]

\[
0 = \frac{\partial \mathcal{L}}{\partial \lambda} = -ne^{-\lambda} + \sum_{j=1}^{k} \left( z_j e^{-\lambda p_j} - (n - z_j) \right).
\]

By adding and subtracting terms of \( (p_j\mu)/\lambda \), (4) may be rewritten as

\[
0 = \frac{\partial \mathcal{L}}{\partial \lambda} = -ne^{-\lambda} + \sum_{j=1}^{k} \left( \frac{z_j e^{-\lambda p_j}}{1 - e^{-\lambda p_j}} - (n - z_j) \right) + \sum_{j=1}^{k} \frac{p_j \mu}{\lambda},
\]

In (5), the part in braces is equal to \((1/\lambda)(\partial \mathcal{L}/\partial p_j)\), which is equal to 0. So, we may solve for \( \mu \):

\[
0 = \frac{-ne^{-\lambda}}{1 - e^{-\lambda}} - \frac{\mu}{\lambda}
\]

\[
\mu = \frac{-\lambda ne^{-\lambda}}{1 - e^{-\lambda}}
\]
By substituting this value for \( \mu \) into (3), the MLE for \( p_i \) is found to be
\[
\hat{p}_i = -\frac{1}{\lambda} \log \left( \frac{n_i(1 - e^{-\lambda}) - z_i}{n_i(1 - e^{-\lambda})} \right), \quad j = 1, \ldots, k,
\]
and the MLE for \( \lambda \) follows from the fact that the \( p_i \) sum to 1:
\[
\hat{\lambda} = -\sum_{j=1}^{k} \log \left( \frac{n_i(1 - e^{-\lambda}) - z_i}{n_i(1 - e^{-\lambda})} \right).
\]
The system of (6) and (7) cannot be solved explicitly, but it is straightforward to find \( \hat{p} \) and \( \hat{\lambda} \) iteratively to arbitrary precision. For reasons that become clear shortly, we call these estimators the partial sample MLEs.

**ESTIMATION FROM THE WHOLE SAMPLE OF HOSTS**

The quantity \( n_i(1 - e^{-\lambda}) \) in Equation 7 can be seen to estimate the total number of hosts sampled, both infected and uninfected. In fact, it is possible to improve the estimation of \( p \) and \( \lambda \) by including information, which should typically be available, on the number of uninfected hosts sampled. Since Rannala et al. (2000) use only the infected hosts, their log-likelihood (1) contains the term \( -n \log(1 - e^{-\lambda}) \), which arises from conditioning upon the hosts being infected. If one uses the information in both infected and uninfected hosts, that term vanishes from the log-likelihood. Doing so, and letting \( M \) denote the total number of hosts (infected and uninfected) sampled, the log-likelihood becomes
\[
\ell = \sum_{i=1}^{k} \left[ z_i \log(1 - e^{-\lambda p_i}) - (M - z_i)\lambda p_i \right].
\]
Since the \( p_i \) and \( \lambda \) occur together only as the products \( \lambda p_i \) in this expression, the unconstrained maximization procedure pursued by Rannala et al. would be appropriate, and it follows that the MLEs (which we call the complete sample MLEs) are
\[
\hat{p}_i = -\frac{1}{\lambda} \log \left( \frac{M - z_i}{M} \right), \quad j = 1, \ldots, k \quad (8)
\]
\[
\hat{\lambda} = -\sum_{j=1}^{k} \log \left( \frac{M - z_i}{M} \right) \quad (9)
\]
These expressions are identical to those in Rannala et al. (2000) except that \( M \) is used in place of \( n \).

Another, more intuitive, derivation of the complete sample MLEs can be given: Let \( 1 - q_j \) be the probability that a host carries no bacteria of allelic type \( j \). Then, detecting the presence or absence of allelic type \( j \) in a host is a Bernoulli trial with probability of success (i.e., probability of presence of the allele) \( q_j \). Since each host sampled is an independent Bernoulli trial, the number of hosts infected with allelic type \( j \) from a total sample size of \( M \) is binomially distributed, and the MLE of \( q_j \) is just the familiar MLE for a binomial proportion, \( \hat{q}_j = z_j/M \). By the assumptions of the infection model, \( 1 - q_j \) is given by the zero term of the Poisson distribution, \( e^{-\lambda p_j} \). So, by the invariance of MLEs to transformation, the MLE of the product, \( \lambda p_j \), can be found by solving \( \hat{q}_j = 1 - e^{-\lambda \hat{p}_j} \). Doing so and then imposing the constraint that \( \sum_{j=1}^{k} p_j = 1 \) gives the estimators in (8) and (9).

If \( M \), the total number of hosts sampled, is available, (8) and (9) are the preferred estimators for two reasons. First, being available in closed form, it is not necessary to iteratively compute the complete sample MLEs as it is for the partial sample MLEs. And second, it can be shown that if \( M \) is known, then \( \lambda \) and \( p \) are minimal sufficient statistics for the parameters \( \lambda \) and \( p \). Since \( M \) is not a deterministic function of \( n \) it follows from the definition of minimal sufficiency that \( n \) and the set of \( z_j \) (\( j = 1, \ldots, k \)) are the minimal sufficient statistics for the parameters \( \lambda \) and \( p \). Since \( M \) is not based on all the available information unless the investigator failed to record or does not actually know \( M \). Since estimators based on sufficient statistics typically have smaller variance than those that are not based on sufficient statistics, the complete sample MLE should be used when \( M \) is known. Of course, if only \( n \) and the \( z_j \)'s are known, then, in that context, the partial sample MLE is based on the sufficient statistic and should be used. Simulations (Figure 2) confirm that the complete sample MLE has lower variance than the partial sample MLE.

When \( \lambda \) is large, so that most of the sampled hosts are infected, the complete sample and partial sample MLEs will differ little from the estimator derived by Rannala et al. (2000). They will differ more from the estimator in Rannala et al. (2000) when \( \lambda \) is small and many of the hosts sampled are not infected by any bacteria. We have empirically found that the MLEs we derive usually give estimates that are between the estimates from Rannala et al.'s estimator and the "uncorrected estimates" of \( z_i/\sum_{i=1}^{k} z_i \). Our MLEs do not fall between the other two estimators only at alleles for which all the estimators give very similar frequency estimates. Hence, while the uncorrected estimator used in Wang et al. (1999) and Qu et al. (1997) tends to overestimate low-frequency alleles and underestimate high-frequency alleles, the estimator proposed by Rannala et al. (2000) overcorrects the problem, underestimating low-frequency alleles and overestimating high-frequency alleles. As our simulations show, the estimators we propose don't suffer from this bias.

Programs to compute the partial sample and complete sample MLEs described above may be downloaded from http://www.rannala.org.

**SIMULATIONS AND DATA**

We repeated a small set of simulations like those carried out in Rannala et al. (2000). Briefly, hosts were
Letter to the Editor

Figure 1.—Figures from simulations like those in Rannala et al. (2000) for the case $\lambda = 2$. R indicates the estimator from Rannala et al. P denotes the partial sample MLE (Equations 6 and 7), and C denotes the complete sample MLE (Equations 8 and 9). We show results for just one sample size, $n = 100$ infected hosts. For the C estimates, the additional uninfected hosts are used in the estimation. (a) Mean values of the estimates for $\lambda$ from data on a three-allele locus with $p_1 = p_2 = 0.5(1 - p_1)$ based on 50,000 simulated replicates. The R estimator consistently overestimates $\lambda$, while both the partial and complete sample estimators estimate $\lambda$ well (though the partial sample MLE tends to overestimate $\lambda$ when the allele frequencies are highly uneven). (b) Percentage of standardized bias for the estimate of $p_1$ in the three-allele locus described above. The arching curve is the same found in Figure 1 of Rannala et al. (2000). The other two curves show that the estimators we propose estimate $p_1$ with very little bias.

We also reanalyzed the tick data given in Table 1 of Rannala et al. (2000), using the partial sample MLE. Our estimates of the allele frequencies for all of the alleles at ospA and all of the alleles except allele D at ospC were intermediate to the “uncorrected estimate” and the Rannala et al. estimate. The three different estimates for allele D at ospC differed only at the fourth decimal place. Our method estimates that the 

*Borrelia burgdorferi* allele frequencies are somewhat more uniform than inferred by Rannala et al. (2000).

Though the model presented by Rannala et al. (2000) was used only in estimating allele frequencies of Borrelia found in ticks, it may be useful in other host/parasite systems. Further, by defining “hosts” to be samples of DNA pooled from different sources, the model, or some altered version of it, may be useful in other situations involving the estimation of allele frequencies from presence/absence data.

We thank Bruce Rannala for comments on a draft of this letter and the members of the statistical genetics seminar group in the Departments of Statistics and Biostatistics at the University of Washington for helpful discussions, particularly Elizabeth Thompson, Matthew...
Stephens, Andrew George, and Anne-Louise Leutenegger. We also thank two anonymous referees for helpful and insightful comments. E.C.A. was supported by National Science Foundation grant BIR-9807747 and P.A.S. was supported by National Institutes of Health grant T32 HG00035-06.

LITERATURE CITED


Communicating editor: G. B. Golding